

ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΠΤΥΞΗ ΠΟΛΥΓΛΩΣΣΗΣ ΟΡΟΛΟΓΙΚΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΟΝ ΓΝΩΣΤΙΚΟ ΤΟΜΕΑ ΤΗΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

Άλκηστις Χιδίρογλου, Παναγιώτης Αρβανίτης, Παναγιώτης Παναγιωτίδης

ΠΕΡΙΛΗΨΗ

Η ανακοίνωση αυτή έχει στόχο να περιγράψει τον σχεδιασμό και την δημιουργία ενός πολύγλωσσου λεξικού όρων που αναφέρονται στο γνωστικό πεδίο της γενικής και εφαρμοσμένης γλωσσολογίας.

Ειδικότερα αναλύονται οι βασικές μεθοδολογικές αρχές που εφαρμόστηκαν, περιγράφεται η ορολογική βάση δεδομένων που δημιουργήθηκε για την συλλογή και επεξεργασία των γλωσσολογικών όρων και αναλύονται ζητήματα σχετικά με την επιλογή τους, την σημασιολογική τους επεξεργασία και την τυποποίησή τους.

Περιγράφεται επίσης ο σχεδιασμός των οθονών διεπαφής (interfaces) χρήστη - Βάσης Δεδομένων και οι τρόποι πρόσβασης των χρηστών σ' αυτήν, εφόσον απώτερος στόχος είναι η δημιουργία ενός πλήρους ηλεκτρονικού περιβάλλοντος αναζήτησης, επισήμανσης και ανάκτησης γλωσσολογικών όρων για το διαδίκτυο. Η ανακοίνωση συνοδεύεται από παρουσίαση της Βάσης σε υπολογιστή.

DESIGN AND DEVELOPMENT OF A MULTILINGUAL TERMINOLOGICAL DATABASE FOR THE SUBJECT FIELD OF APPLIED LINGUISTICS

Άλκηστις Χιδίρογλου, Παναγιώτης Αρβανίτης, Παναγιώτης Παναγιωτίδης

SUMMARY

This paper focuses on the planning and the creation of a multilingual terminological database that refers to the cognitive field of applied linguistics.

More specifically, we analyze the methodological approaches and principles that were used, we describe the database which was created for the collection and processing of the linguistic terms and finally we analyze questions regarding their choice, their semantic processing and their standardization.

Furthermore, the planning of user interfaces is described as well as the different users' access ways since the ultimate goal is the creation of a complete electronic environment of searching and retrieving in the Internet. The paper is accompanied by an electronic presentation.

0. ΕΙΣΑΓΩΓΗ

Σε ένα κοινωνικό περιβάλλον που γίνεται ολοένα και περισσότερο πολύγλωσσο, έχει διαπιστωθεί από καιρό η ανάγκη για την οργανωμένη ανάπτυξη της ελληνικής ορολογίας [1]. Έτσι έχει διαφανεί πως απαιτείται συντονισμός σε πολλά επίπεδα και ένωση δυνάμεων όχι μόνο για την παραγωγή όρων, τον έλεγχο ποιότητας, την πιστοποίηση και την διάδοσή τους, αλλά και για την δημιουργία μιας ισχυρής ηλεκτρονικής υποδομής, με δημιουργία ειδικευμένων ορολογικών βάσεων δεδομένων, ειδικευμένων λεξικών υπό μορφή CD-ROM καθώς και ορολογικών θησαυρών.

Ένα επιστημονικό πεδίο που παρά την εξέλιξη που γνωρίζει τα τελευταία χρόνια έχει λίγα σημεία ηλεκτρονικής αναφοράς και πρόσβασης για την ανεύρεση όρων που θα διευκόλυναν τόσο τους απλούς χρήστες όσο και τους φοιτητές και τους επιστήμονες του χώρου είναι αυτό της Θεωρητικής και της Εφαρμοσμένης Γλωσσολογίας. Με έναυσμα την σκέψη αυτή, η παρούσα ανακοίνωση δεν θα ασχοληθεί με γενικές θεωρητικές αρχές της Ορολογίας αλλά θα επικεντρωθεί στα προβλήματα που προέκυψαν κατά τον σχεδιασμό και την κατασκευή μιας πολύγλωσσης ορολογικής βάσης δεδομένων για τα παραπάνω γνωστικά πεδία.

1. ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΟΡΟΛΟΓΙΑ

Όπως είναι γνωστό η τεχνολογία των Βάσεων Δεδομένων προσφέρει πάρα πολλά τόσο στο χώρο της λεξικογραφίας όσο και σ' αυτόν της ορολογίας, παρέχοντας συστήματα που επιτρέπουν την διαχείριση, αποθήκευση και αναζήτηση μεγάλου αριθμού γλωσσικών δεδομένων.

Σήμερα αν και η ανάγκη εργαλείων διαχείρισης και αποθήκευσης γλωσσικών δεδομένων είναι αυταπόδεικτη και παρά τις τεράστιες συλλογές δεδομένων που βρίσκονται αποθηκευμένες (κυρίως μεγάλα τεχνικά και διοικητικά κείμενα, παράλληλα μονόγλωσσα ή πολύγλωσσα κείμενα, φωνητικά δεδομένα, φωνητικές καταγραφές, μορφολογικά και συντακτικά σχήματα, κλπ), παραμένουν εκκρεμή πολλά ζητήματα προς διερεύνηση και γίνεται επιτακτικότερη η αναζήτηση νέων μοντέλων Βάσεων Δεδομένων που θα ήταν γλωσσολογικά περισσότερο κατάλληλες και ευέλικτες. Για παράδειγμα οι πολύγλωσσες Λεξιλογικές και Ορολογικές Βάσεις Δεδομένων που συναντώνται στα Συστήματα υποβοηθούμενης από Μηχανή Μετάφρασης (machine-aided human translation - MAHT) συχνά δεν είναι τίποτε άλλο παρά τεράστιες συλλογές δίγλωσσων λεξικών που αποθηκεύονται με την μορφή πινάκων. Έτσι η ποιότητα της μετάφρασης στηρίζεται πρωτίστως στον όγκο των ειδικευμένων λεξικών που υποστηρίζουν το σύστημα, συνήθως

δεν ενσωματώνονται σ' αυτό και συχνά δέχονται συγκρουόμενες μεταξύ τους μεταφράσεις όρων από εκατοντάδες διαφορετικούς μεταφραστές - χρήστες.

Τα στηριζόμενα στο Σχεσιακό μοντέλο [2], συστήματα Διαχείρισης Βάσης Δεδομένων, που χρησιμοποιούνται ευρέως σήμερα, επιτρέπουν την αποθήκευση γλωσσικών δεδομένων σε αλληλοσυνδεόμενους μεταξύ τους δισδιάστατους πίνακες χρησιμοποιώντας όμως τεράστια υπολογιστική ισχύ που καταναλώνεται για να δεικτοδοτεί λέξεις και όρους στους πίνακες αυτούς, να τις αναζητά και να τις ανακαλεί.

Στην ανάγκη μείωσης του εύρους και της πολυπλοκότητας αυτής προσπάθησαν να απαντήσουν τα λεγόμενα "σημασιολογικά" μοντέλα Βάσεων Δεδομένων, τα οποία άρχισαν να αναπτύσσονται στις αρχές της δεκαετίας του '80, για να καθιερωθούν αργότερα ως "Αντικειμενοστρεφή Συστήματα Βάσεων Δεδομένων" (Object Oriented Database Management Systems - OODBMS) [2]. Στα συστήματα αυτά, τα διαφορετικού είδους δεδομένα αντιμετωπίζονται ως "αντικείμενα" (objects), με συγκεκριμένα χαρακτηριστικά, συγκεκριμένες ιδιότητες και συγκεκριμένες σχέσεις.

Ωστόσο, φαίνεται ότι η ανάπτυξή τους περιορίζεται ακόμη, τόσο από την ύπαρξη διαφορετικών μεθοδολογικών προσεγγίσεων όσο και από την απουσία ισχυρών προτύπων [3].

2. ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

Η προσέγγιση που επιλέχθηκε για την παρούσα Βάση Δεδομένων είναι υβριδική. Για την καταγραφή και συγκέντρωση των όρων προτιμήθηκε ο εξαρχής σχεδιασμός ενός ενοποιημένου περιβάλλοντος διαχείρισης που υλοποιήθηκε σε ένα Σύστημα Σχεσιακής Βάσης Δεδομένων. Η προσπάθεια για την υλοποίηση μιας ορολογικής Βάσης Δεδομένων ξεκινά ήδη από μια αρχική επισήμανση που πρέπει να λαμβάνεται υπόψη κατά το στάδιο του σχεδιασμού της. Σύμφωνα με αυτήν οι ορολόγοι επιστήμονες ενδιαφέρονται πρωτίστως για την έννοια και την απόδοσή της ενώ οι λεξικογράφοι ξεκινούν από το λήμμα για να καταλήξουν στην έννοια.

Ο σχεδιασμός μιας δίγλωσσης, πολύ δε περισσότερο πολύγλωσσης βάσης, αναδεικνύει ζητήματα που αναφέρονται τόσο στο είδος των γλωσσολογικών πληροφοριών που πρέπει να αποθηκευτούν (γραμματικό – μορφοσυντακτικές), όσο και στον τρόπο με τον οποίο πρέπει αυτές οι πληροφορίες να καταγραφούν και να ταξινομηθούν. Ένα ακόμα σημαντικό ζήτημα που πρέπει να ληφθεί υπόψη είναι η αυστηρή τυποποίηση και κατηγοριοποίηση των συγκεντρωμένων δεδομένων, η εναρμόνισή τους -όσο είναι δυνατόν- με διεθνή πρότυπα,

έτσι ώστε να είναι δυνατή η μορφοποιημένη εξαγωγή τους και η ανταλλαγή τους με παρόμοια συστήματα.

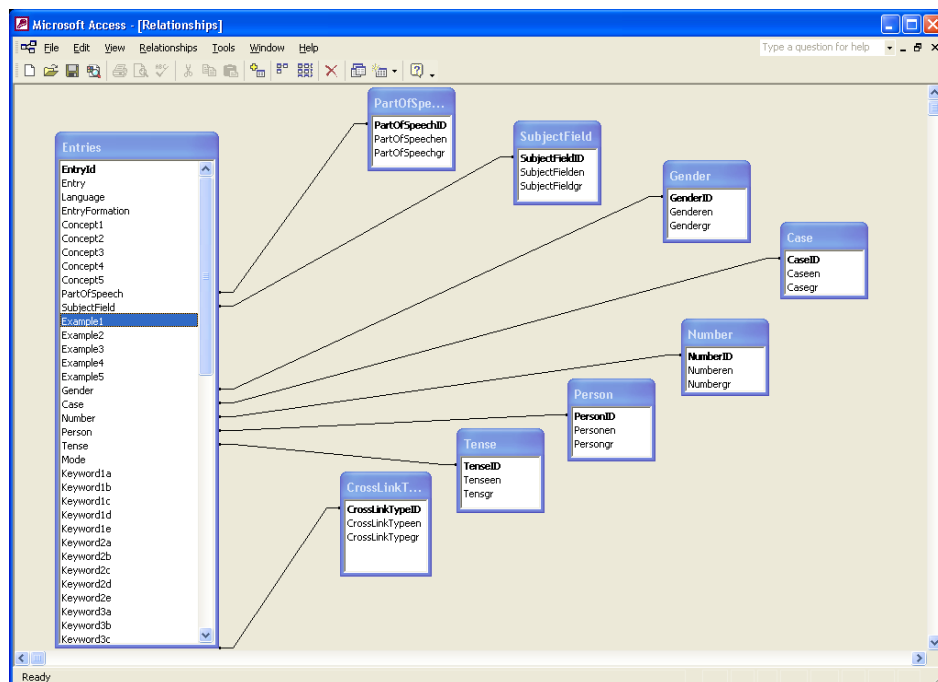
Για την αποφυγή μέρους των παραπάνω προβλημάτων η συγκέντρωση των όρων βασίστηκε αρχικά στην επισκόπηση υλικού το οποίο περιλαμβάνει την ελληνική και ξένη βιβλιογραφία ειδικών λεξικών του συγκεκριμένου γνωστικού πεδίου [4], με πρόθεση αυτή να επεκταθεί αργότερα σε ειδικά περιοδικά, πανεπιστημιακά εγχειρίδια, βιβλία [5], [6], πρακτικά συνεδρίων εφαρμοσμένης γλωσσολογίας, κ.ά.

Κατά τον σχεδιασμό των πεδίων χρησιμοποιήθηκαν οι γενικές οδηγίες τυποποίησης για την ανάπτυξη μονόγλωσσών και πολύγλωσσων θησαυρών [7], [8], και λήφθηκαν εκτενώς υπόψη οι προτεινόμενες κατηγοριοποιήσεις των προτύπων OLIF [9], MARTIF [10]. Τα συγκεκριμένα πρότυπα προτείνουν ένα σύνολο γενικών και ειδικών κατηγοριών και ένα μεγάλο αριθμό επιμέρους χαρακτηριστικών. Απ' αυτά επιλέχθηκαν περίπου 100 και όπου ήταν δυνατόν, η ονοματολογία τους κρατήθηκε ίδια για λόγους συμβατότητας.

3. ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Αναφέρθηκε προηγουμένα πως ο σχεδιασμός της παρούσας Βάσης Δεδομένων για την διαχείριση των ορολογικών δεδομένων έγινε χρησιμοποιώντας τις αρχές των Συστημάτων Σχεσιακών Βάσεων Δεδομένων.

Τμήμα του αρχικού σχεδιασμού έγινε ακολουθώντας μια αντικειμενοστρεφή προσέγγιση, η έλλειψη όμως αυστηρής τυποποίησης σε περιβάλλοντα ανάπτυξης αντικειμενοστρεφών Βάσεων Δεδομένων, επέβαλλε την τελική υλοποίηση σε σχεσιακό μοντέλο. Ειδικότερα ο αρχικός σχεδιασμός έγινε σε Microsoft Access και περιελάμβανε πεδία όρων και εννοιών για έξι γλώσσες (ελληνικά, αγγλικά, γαλλικά, γερμανικά, ιταλικά, ισπανικά). Στη συνέχεια αναπτύχθηκαν δοκιμαστικά περιβάλλοντα σε SQL [11] και σε γλώσσα PHP και MySQL για χρήση στο διαδίκτυο.



ΣΧΗΜΑ 1

Όπως φαίνεται και στο παραπάνω σχήμα 1, που απεικονίζει μερικώς τις σχέσεις μεταξύ πινάκων και πεδίων, η Βάση Δεδομένων περιλαμβάνει 14 πίνακες που αποτελούνται από ένα σύνολο 180 πεδίων περίπου που καλύπτουν:

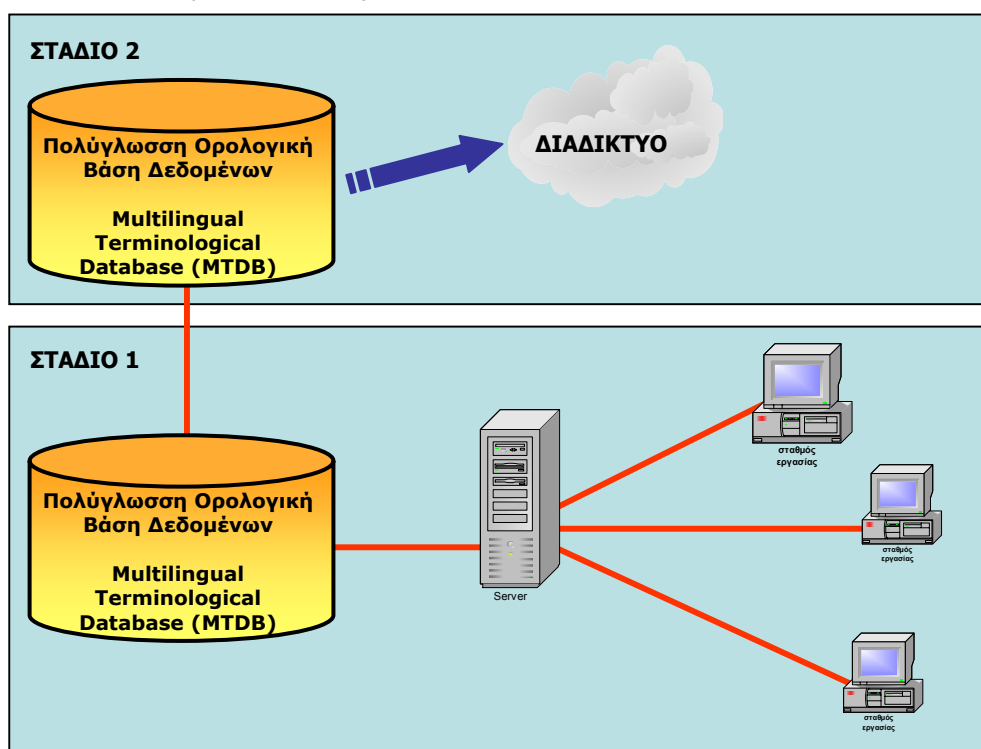
- Ορισμό όρου
- Συνώνυμα / Αντώνυμα όρου
- Σχετιζόμενους όρους
- Θεματικό πεδίο / υποπεδίο
- Μορφολογικές πληροφορίες
- Γραμματικές πληροφορίες (μέρος λόγου, κλπ)
- Παραδείγματα χρήσης
- Δημιουργό όρου (αν υπάρχει)
- Πηγή όρου (με πλήρη αναφορά σε βιβλιογραφία αν υπάρχει)
- Διάφορες τεχνικές πληροφορίες (καταγραφείας όρου, ημερομηνία εισαγωγής όρου, ημερομηνία ελέγχου και επικύρωσης, κ.ά)

- Συγκείμενο όρου
- Λέξεις-κλειδιά ανεύρεσης όρου
- Παρατηρήσεις και γενικά σχόλια

Η καταγραφή αποφασίστηκε να γίνει σε δύο στάδια. Στο πρώτο, μικρές ομάδες ερευνητών – σπουδαστών ασχολούνται με την εύρεση, επισκόπηση και καταγραφή των όρων ενώ στην δεύτερη μια ολιγομελής ομάδα 10 μεταπτυχιακών φοιτητών ελέγχει, αξιολογεί και «επικυρώνει» τα ήδη συγκεντρωμένα δεδομένα.

4. ΗΛΕΚΤΡΟΝΙΚΟ ΠΕΡΙΒΑΛΛΟΝ ΔΙΑΧΕΙΡΙΣΗΣ

Η παρούσα Βάση Δεδομένων αρχικά σχεδιάστηκε και υλοποιήθηκε για χρήση μόνο σε περιβάλλον τοπικού δικτύου Η/Υ. Όπως φαίνεται και στο σχήμα 2, κατά το πρώτο στάδιο, ένας κεντρικός υπολογιστής διανέμει την Βάση σε 20 τερματικούς σταθμούς εργασίας, όπου γίνεται εισαγωγή όρων από ισάριθμους καταγραφείς. Η αποθηκευμένη Βάση Δεδομένων είναι προσβάσιμη μόνο σε αυτούς.



ΣΧΗΜΑ 2

Στην συνέχεια, στο δεύτερο στάδιο, τα περιεχόμενα της Βάσης ελέγχονται και «επικυρώνονται» από μια μικρή ομάδα ερευνητών και κατόπιν μπορούν να διανεμηθούν ελεύθερα στο διαδίκτυο.

Το υπόλοιπο τμήμα της Βάσης Δεδομένων αποτελείται από τις οθόνες διεπαφής χρήστη – Βάσης για την εισαγωγή και διαχείριση των όρων. Σύμφωνα με τον σχεδιασμό, η Βάση Δεδομένων στο το πρώτο στάδιο επιτρέπει την:

- εισαγωγή νέων όρων,
- διόρθωση όρων,
- επιλεκτική διαγραφή όρων ή πληροφοριών αυτών.

Ενώ στο το δεύτερο στάδιο επιτρέπει την:

- διόρθωση και επιλεκτική διαγραφή όρων ή πληροφοριών αυτών,
- αναζήτηση όρων από ανεξάρτητους χρήστες (για χρήση στο Διαδίκτυο).

5. ΕΠΙΛΟΓΟΣ

Είναι προφανές πως η προσπάθεια συγκέντρωσης και απόδοσης όρων που αναφέρονται στο δυναμικά εξελισσόμενο χώρο της Θεωρητικής και Εφαρμοσμένης Γλωσσολογίας είναι ένα εξαιρετικά δύσκολο εγχείρημα, που απαιτεί μακροχρόνιο σχεδιασμό, συντονισμό σε πολλά επίπεδα και συνεχή ανταλλαγή πληροφοριών και απόψεων. Για τον λόγο αυτό λήφθηκε ιδιαίτερη μνεία, ώστε η αρχιτεκτονική του συστήματος να είναι ανοιχτή, να επιδέχεται συνεχείς διορθώσεις και βελτιώσεις και επίσης έγινε προσπάθεια ώστε οι κατηγοριοποιήσεις των γλωσσικών δεδομένων να υπακούουν όσο το δυνατόν σε διεθνή πρότυπα τυποποίησης [12], [13], [14].

6. ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1]: Καραγιάννης, Γ. (2001), Εθνικό Πρόγραμμα Ορολογικού συντονισμού (ΕΠΟΣ): Στρατηγική μελέτη, μεθοδολογία και στόχοι, Πρακτικά 3^{ου} συνεδρίου Ελληνικής Γλώσσας και ορολογίας, Αθήνα, σελ. 324.
- [2]: Date, C. J. (1995). An Introduction to Database Systems. Massachusetts: Addison-Wesley Publishing Company, ελλ. μτφ. Άλβας, Τάκης (1996), Εισαγωγή στα Συστήματα Βάσεων Δεδομένων. Αθήνα: εκδ. Κλειδάριθμος, τομ. Α', Β'.
- [3]: BLAKOWSKI, Gerold & SAMPATHKUMAR, Srihari (1994). "Multimedia Storage and Databases", *Perspectives of Multimedia Systems: Reports from the 1994 Dagstuhl Multimedia Seminar, Technical Report GIT-GVU-95-23*, pp.31-39, Atlanta: Graphics, Visualisation and Usability Laboratory and College of Computing, Georgia Institute of Technology
Electronic version: <ftp://ftp.gvu.gatech.edu/pub/gvu/tech-reports/95-23.ps.Z>

- [4]: Galisson, R. & Coste, D. (1976). Dictionnaire de Didactique des Langues. Paris: Hachette
- [5]: Τοκατλίδου, Βάσω (1986). Εισαγωγή στην Διδακτική των Ζωντανών Γλωσσών. Προβλήματα-Προτάσεις. Αθήνα: Εκδόσεις Οδυσσέας
- [6]: Τοκατλίδου, Βάσω (2003). Γλώσσα, επικοινωνία και γλωσσική εκπαίδευση. Αθήνα: Εκδόσεις Πατάκη
- [7]: ISO 2788-1986, "Guidelines for the establishment and development of monolingual thesauri – Documentation", International Organization for Standardization, 1986.
- [8]: ISO 5965-1985, "Guidelines for the establishment and development of multilingual thesauri – Documentation", International Organization for Standardization, 1985.
- [9]: Formalization of the Data Categories for the Open Lexicon Interchange Format (OLIF), July 4 2003, OLIF Consortium Web site <http://www.olif.net/>
- [10]: ISO 12200 (1998). Computer applications in terminology - Machine-readable terminology interchange format (MARTIF) -Negociated interchange ISO/DIS 12200, ISO, Genève.
- [11]: ISO/IEC 9075-4:1996 (1996). Information technology - Database languages - SQL - Part 4: Persistent Stored Modules (SQL/PSM), ISO, Genève.
ISO/IEC 9075-3:1995 (1995). Information technology - Database languages - SQL - Part 3: Call-Level Interface (SQL/CLI), ISO, Genève.
ISO/IEC DIS 9075-1 (1999). Information processing systems - Data base language SQL - Part 1: Frame, ISO, Genève
- [12]: Sager, J. C., L'Homme, M.-C. (1994). A model for the definition of concepts: Rules for analytical definitions in terminological databases. Terminology 1(2): 351-374.
- [13]: ISO 1951:1997 (1997). Lexicographical symbols and typographical conventions for use in terminography ISO 1951:1997, ISO, Genève.
- [14]: ISO 12618:1994 (1994). Computational aids in terminology - Creation and use of terminological databases and text corpora ISO/TR 12618:1994, ISO, Genève.

Άλκηστις Χιδίρογλου

Επίκουρη Καθηγήτρια του Τμήματος Γαλλικής Γλώσσας και Φιλολογίας του Α.Π.Θ.,
alkisti@frl.auth.gr

Παναγιώτης Αρβανίτης

Λέκτορας του Τμήματος Γαλλικής Γλώσσας και Φιλολογίας του Α.Π.Θ.,
arva@frl.auth.gr

Παναγιώτης Παναγιωτίδης

Λέκτορας του Τμήματος Γαλλικής Γλώσσας και Φιλολογίας του Α.Π.Θ.,
pana@frl.auth.gr